IBM TRECVID 03 Search System

Arnon Amir, Marco Berg, Matthew Hill, Giri Iyengar, Ching-Yung Lin, Milind Naphade, Apostol (Paul) Natsev, Chalapathy Neti, Harriet Nock, John Smith, Belle Tseng

Nov 18th , 2003

IBM Systems: Overview

Speech-based Retrieval

"White House with running fountain"

Process Query

Rank Shots

Interactive
Manual
Automatic

Fuse

Video-based Retrieval

Process Query

Rank Shots

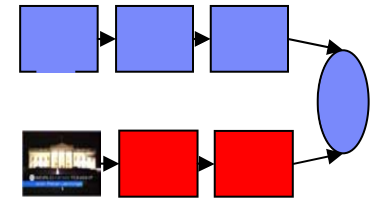
3 different strategies

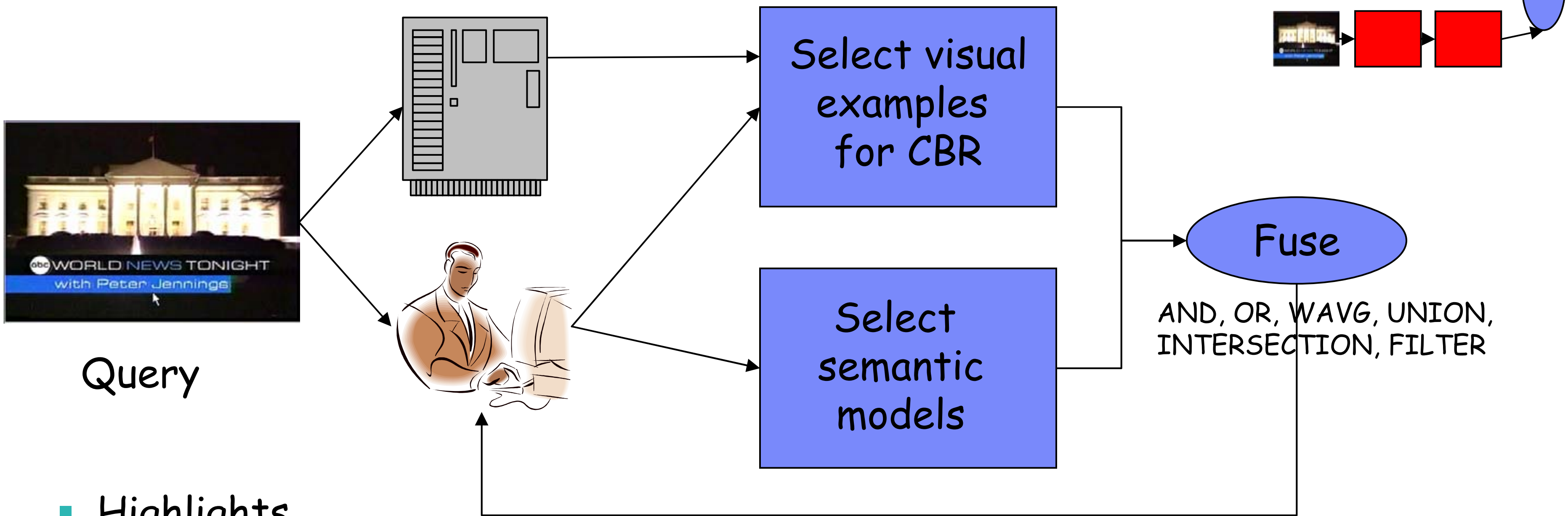
Interactive
Manual
Automatic

Highlights:

- Fusion of two independent unimodal runs
- Fully automatic runs in both modalities

Visual Retrieval System



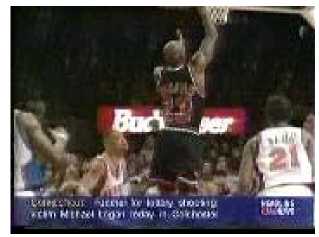
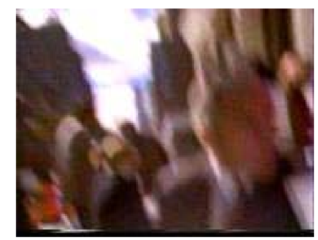
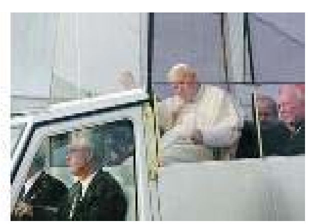


- Highlights
  - Visual features: color, texture, edges, shape, motion, model vectors
  - Semantic features: limited semantic vocabulary (approx. 70 statistical models)
  - Filters: news, commercials, CNN/ABC/C-SPAN, videos, clusters
- Performance (MAP)
  - Interactive CBR/MBR: 0.127
  - Manual CBR/MBR: 0.046
  - Automatic CBR: 0.043

Query Formulation

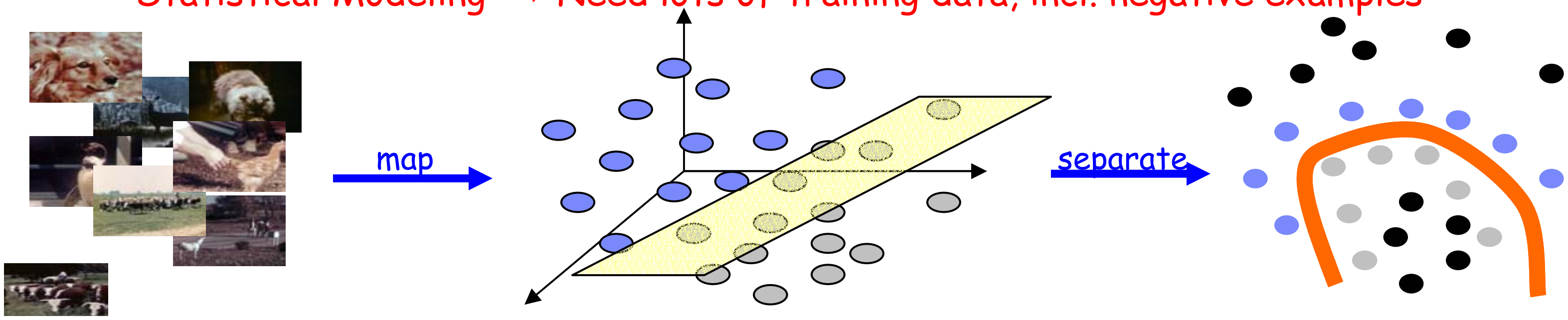
- Textual query formulation
  - Keyword-based
  - Boolean keyword-based
  - Example:
    - Query topic 113: *Find shots with one or more snow-covered mountain peaks or ridges. Some sky must be visible behind them.*
    - Manual keyword query: snow cover mountain peak ridge sky visible
    - Automatic keyword query: Remove "Find shots with one or more" prefix
    - Manual Boolean query: (ski | downhill) & mountain & (snow | glacier | cliff) & (snow-storm) & (summit | peak) & (rocky | himalayas | antarctica | Alaska | everest) & (climbers & rescue | fall | avalanche)
- Visual query formulation
  - Content-based
    - Query with each positive example
    - Use OR semantics for fusing results from multiple queries
  - Model-based
    - Like CBR but using semantic features (model vectors)
    - MBR query 117: 1.0 People – 0.5 Indoors – 0.5 Sport_Event
  - Boolean content-based/model-based?

Visual Query Examples: What Is A Picture Really Worth?

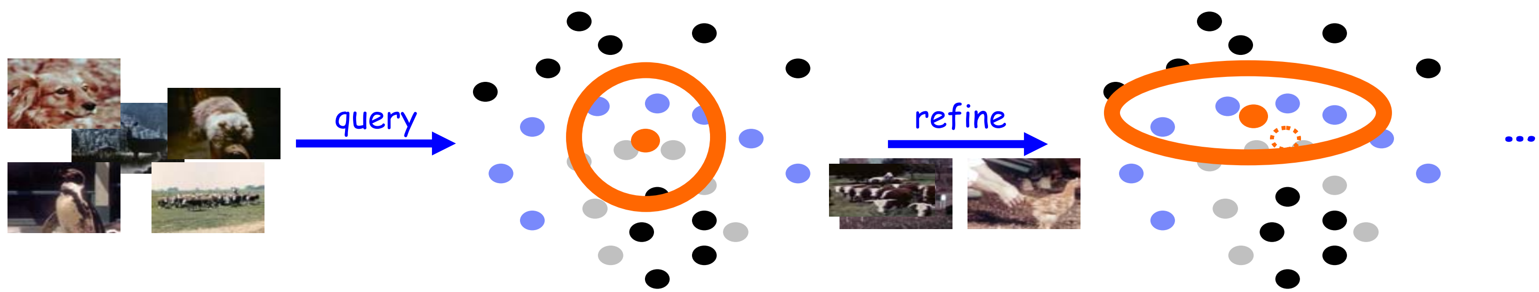
Query Topics	Query Topic Examples		
	The Good	The Bad	The Ugly
Find scene: Aerial views with roads & buildings			
Find event: Basketball score			
Find object: Cup of coffee			
Find person: Pope John Paul II			

Visual Query Formulation: Approaches

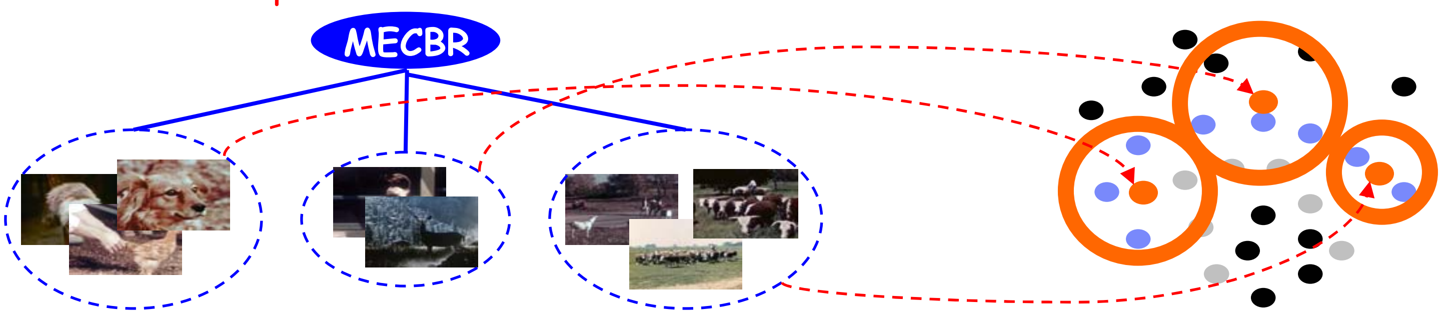
Statistical Modeling → Need lots of training data, incl. negative examples



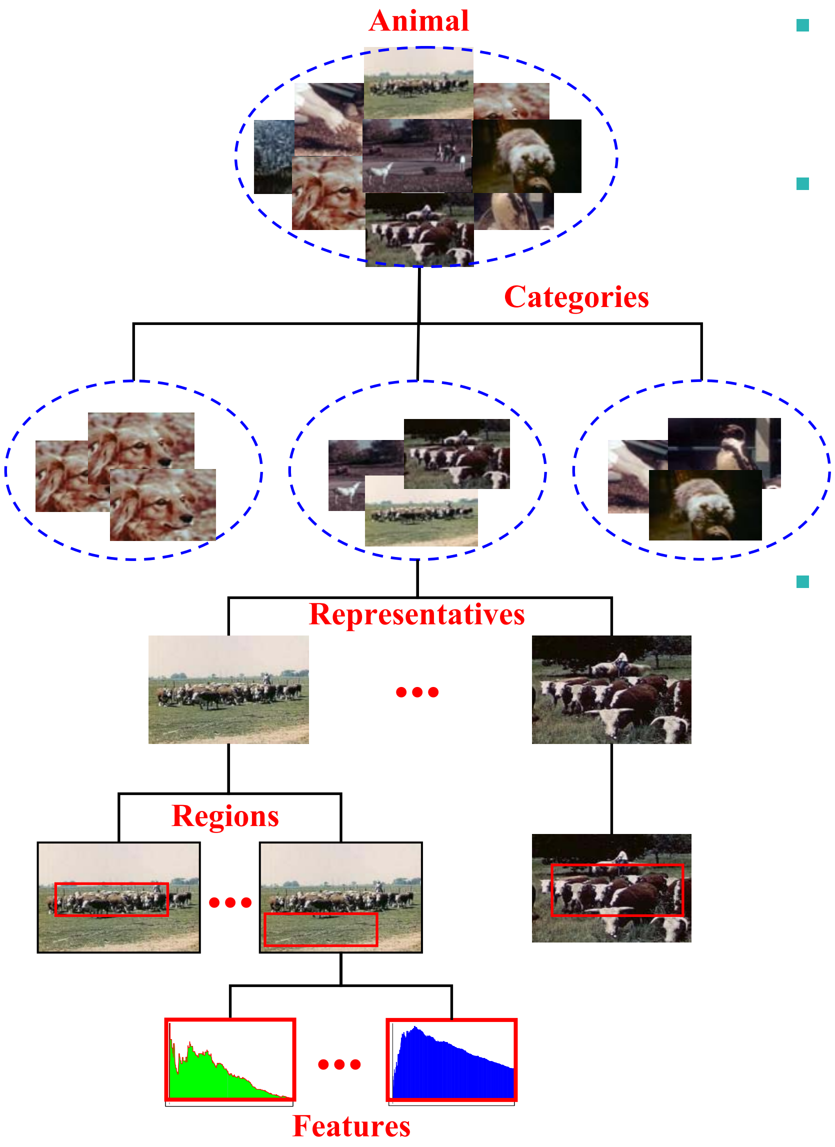
Relevance Feedback → handles rare classes but not diversity; requires interaction



Multi-Example CBR → addresses rare & diverse semantic classes; no interaction



Multi-Example Content-Based Retrieval (MECBR)



- Problem
  - Given a (small) set of concept exemplars, learn concept representation & formulate visual query
- Approach: bridge gap between CBR and statistical modeling
  - Categorize examples into distinct visual subsets
  - Select representative(s) for each category
  - Execute content-based query with each representative
  - Fuse results within/across categories
- Issues
  - Categorization: GMM, clustering, greedy
  - Representatives: centroid, weighted sampling
  - Feature selection: color, texture, edge, models
  - Feature granularity: global, regional (layout, grid)
  - Feature ambiguity: multiple-instance learning
  - Fusion:
    - AND logic within categories
    - OR logic between categories

MECBR Approach Details

- Step 1: Categorize examples:
  - K-means, GMM unreliable (too few examples)
  - Use greedy selection to order & select examples iteratively by their "distinction"
  - Distinction measured as distance to closest previously formed category
  - If distinction > cluster radius threshold, label example as "distinct" (new category)
  - If not, categorize example to closest cluster
- Step 2: Select category representatives
  - Statistical cluster measures not robust (unreliable means, singular variances)
  - Use weighted sampling of category examples
  - Weights proportional to distance of representative to cluster centroid
- Step 3: Execute content-based queries
  - 166-D HSV color correlograms & 46-D model vectors with statistical normalization
  - Query example model vectors automatically tell us which models "fired" up
  - Feature granularity: global for query examples and global/regional for target images
- Step 4: Aggregate content-based retrieval results
  - Feature fusion: similarity score averaging
  - Example fusion (same category): AND logic (weighted AVG of similarity scores)
  - Category fusion: OR logic (MAX similarity)

Visual Categorization Example: Basketball

1.

2.

3.

4.

5.

6.

7.

8.

9.

Automatic Visual Query Formulation: Summary

- Challenges
  - No prior knowledge of query topic, examples, or dataset
  - Unreliable features when using few examples
  - More examples not always good—a single poor example could be devastating
  - Differentiating between good and bad (resolving ambiguity) is not easy…
  - Robust automatic categorization is also hard
- Text processing analogs
  - MECBR -> Boolean text queries
  - Clustering & feature aggregation -> stemming
  - Weighted cluster sampling -> removing stop words
- Some lessons
  - Categorization improves performance by 30-40%
  - Semantic features outperform visual features by 10-15%
  - Regional matching outperforms global matching by 5-10%
  - Fusion of features, examples, and categories boosts performance by 30-50%
  - Automatic MECBR run performs within 10% of manual run!

IBM Systems: Overview

Speech-based Retrieval

"White House with running fountain" → Process Query → Rank Shots → Fuse

Interactive
Manual
Automatic

Video-based Retrieval

→ Process Query → Rank Shots → Fuse

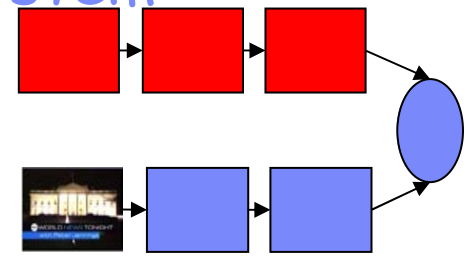
3 different strategies

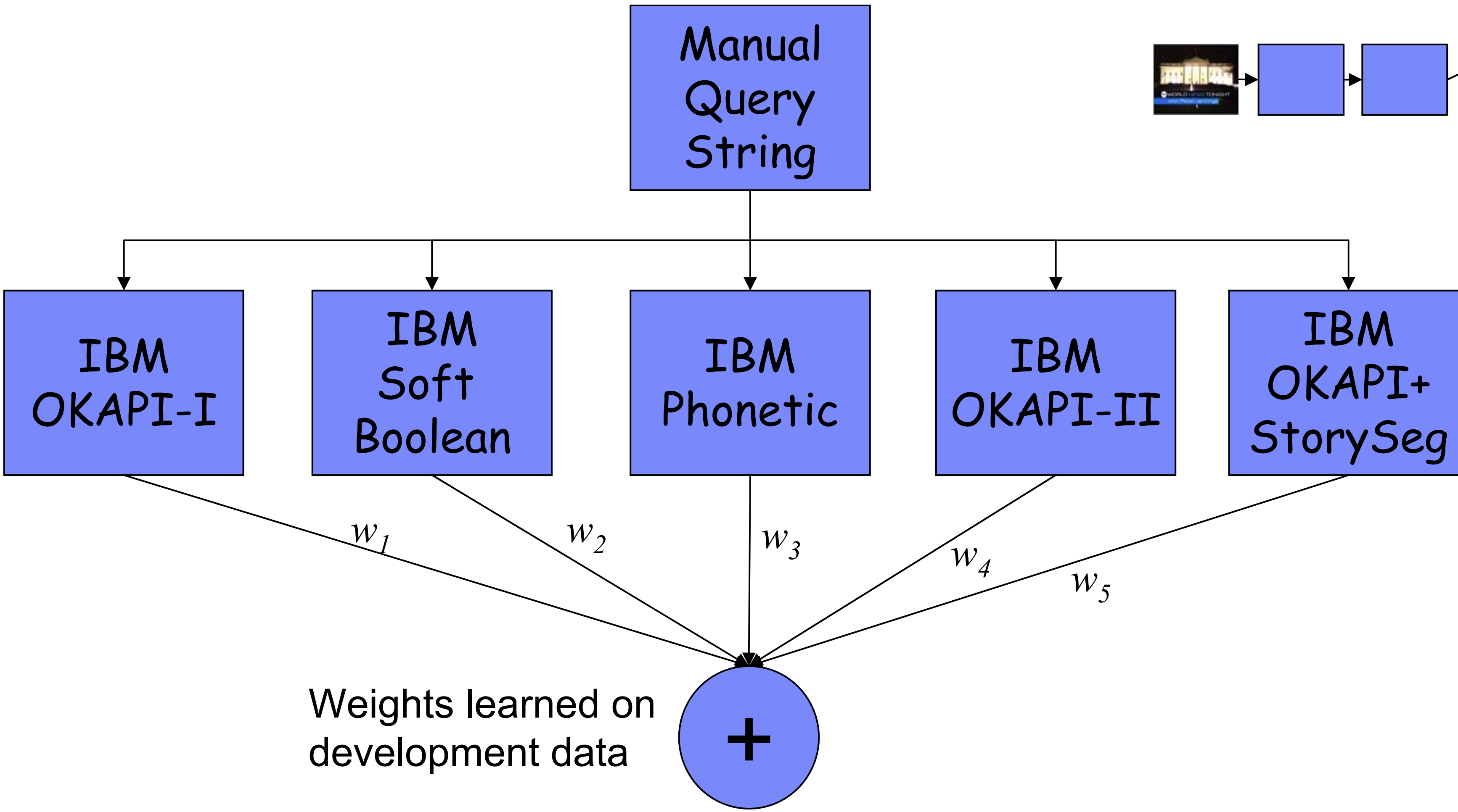
Interactive
Manual
Automatic

Highlights:

- Fusion of two independent unimodal runs
- Fully automatic runs in both modalities

Fusion SDR - Extension of IBM TRECVID02 System

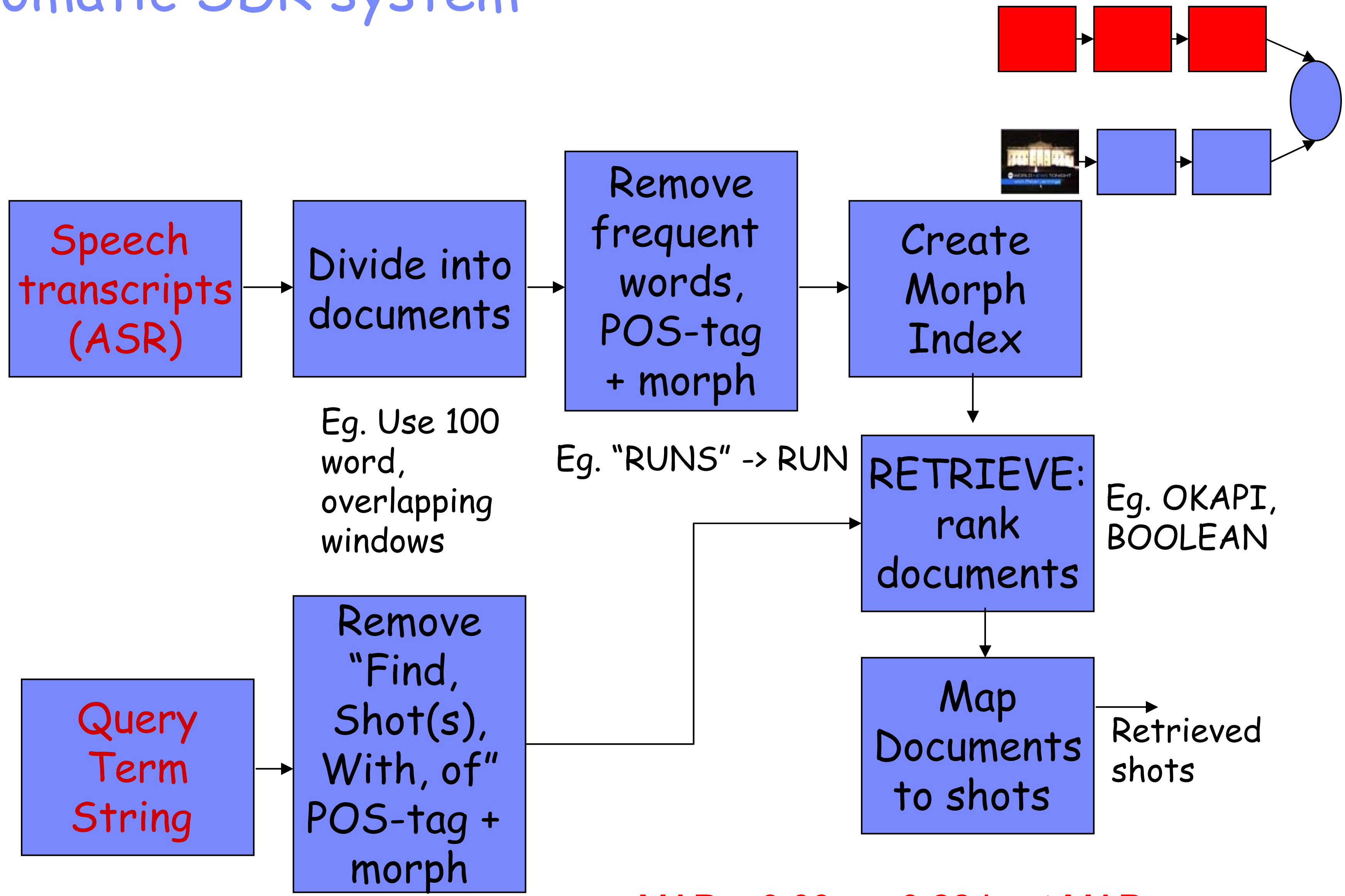




Fusion system performance on dev. set is 25% higher than of best individual system

Best IBM Unimodal system. MAP = 0.12

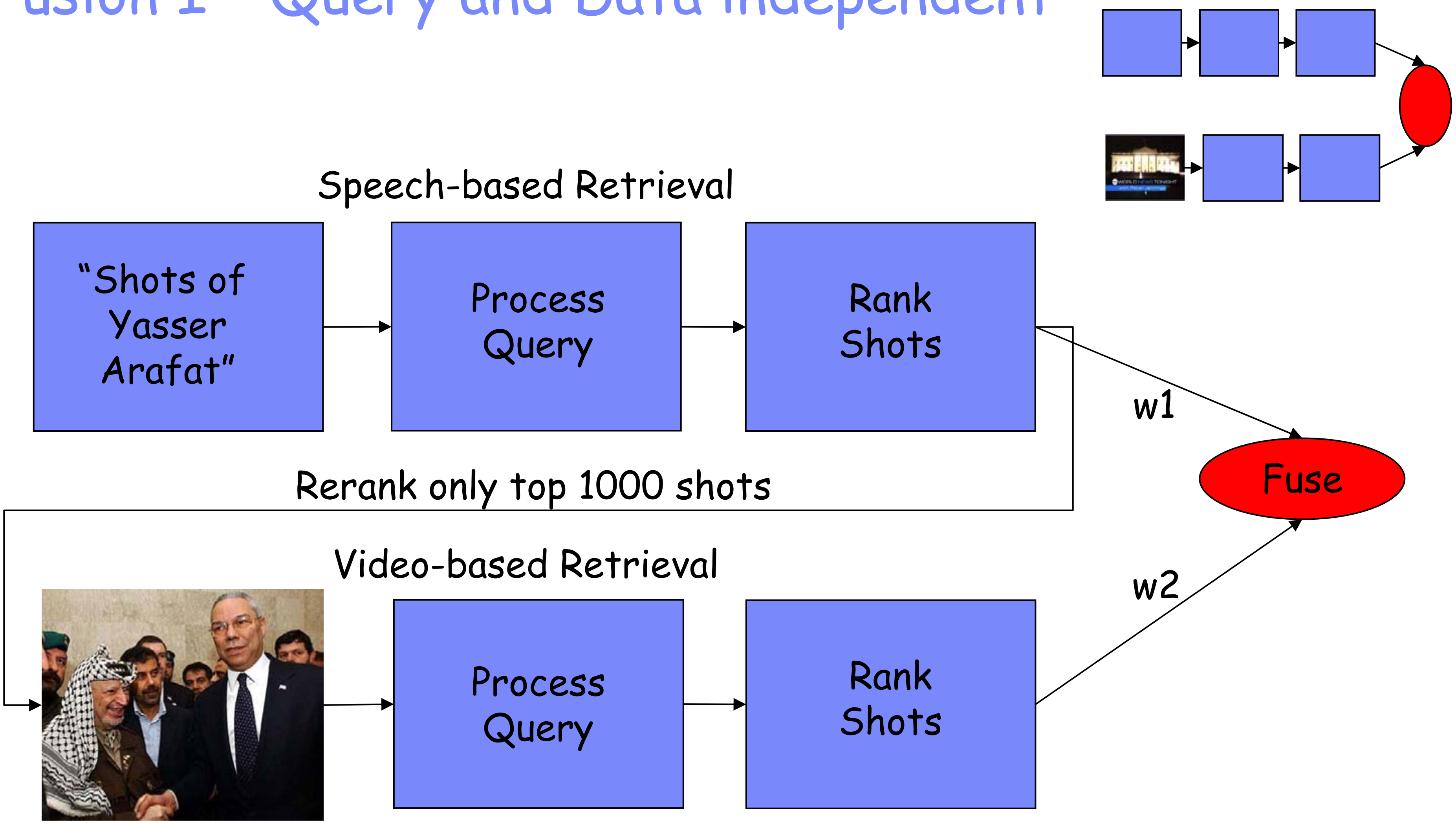
Automatic SDR system



(done automatically)

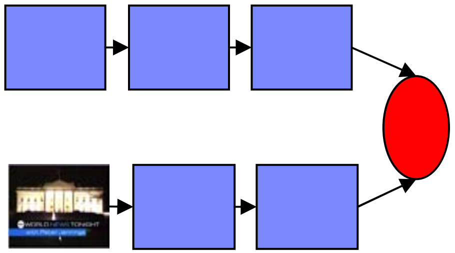
MAP = 0.09 vs. 0.22 best MAP

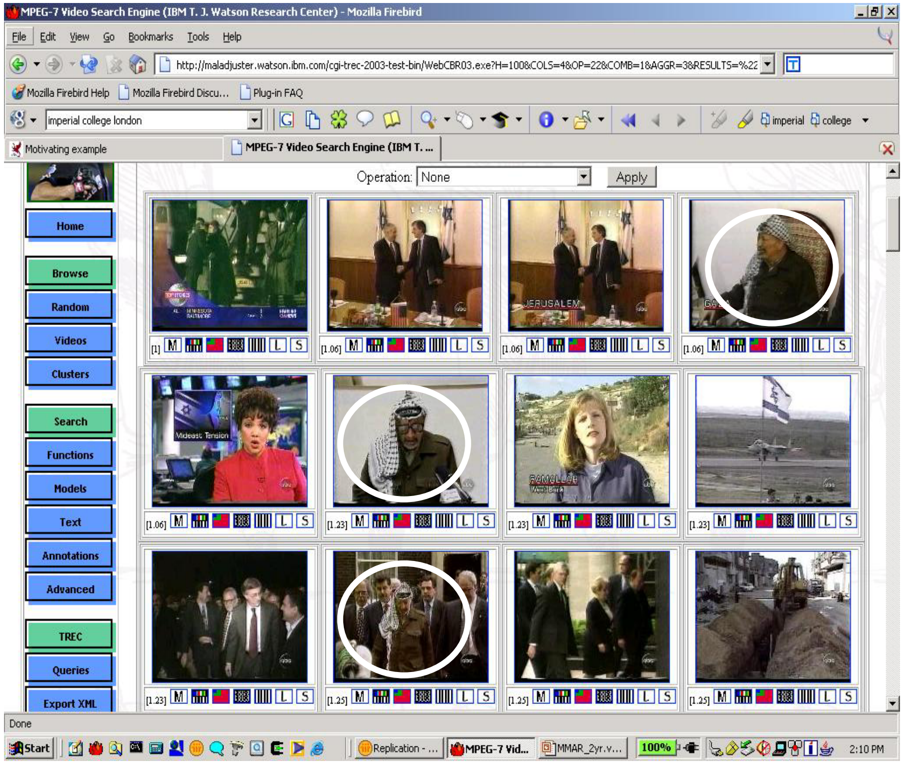
Fusion I - Query and Data independent



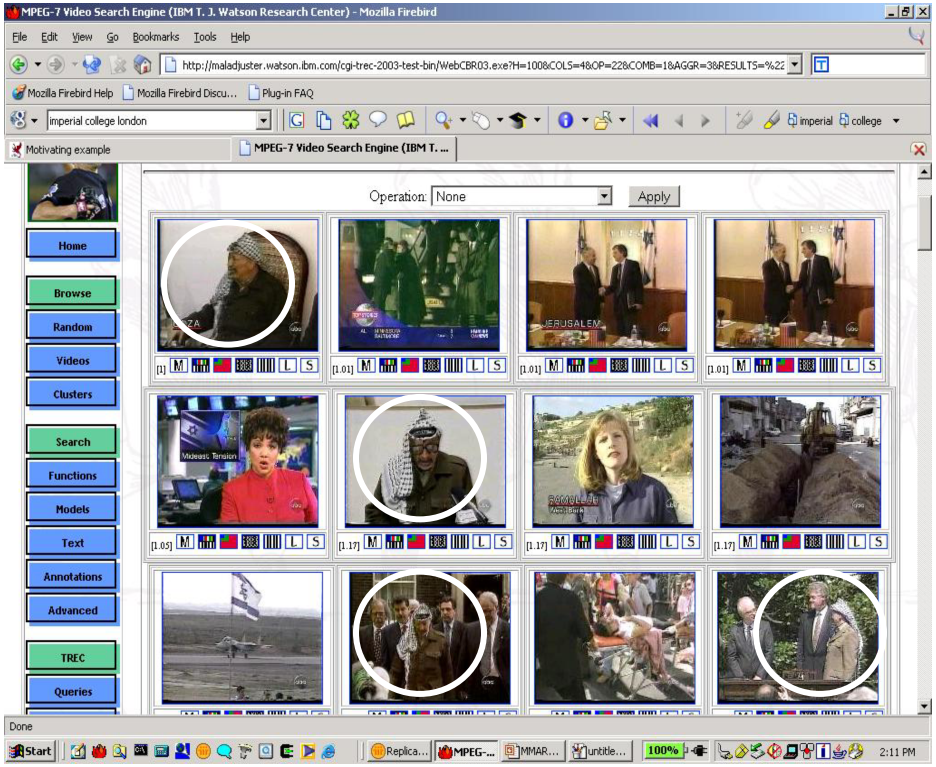
w1 and w2 are query and data independent (hurts?). MAP = 0.123

Fusion I - Query and Data independent



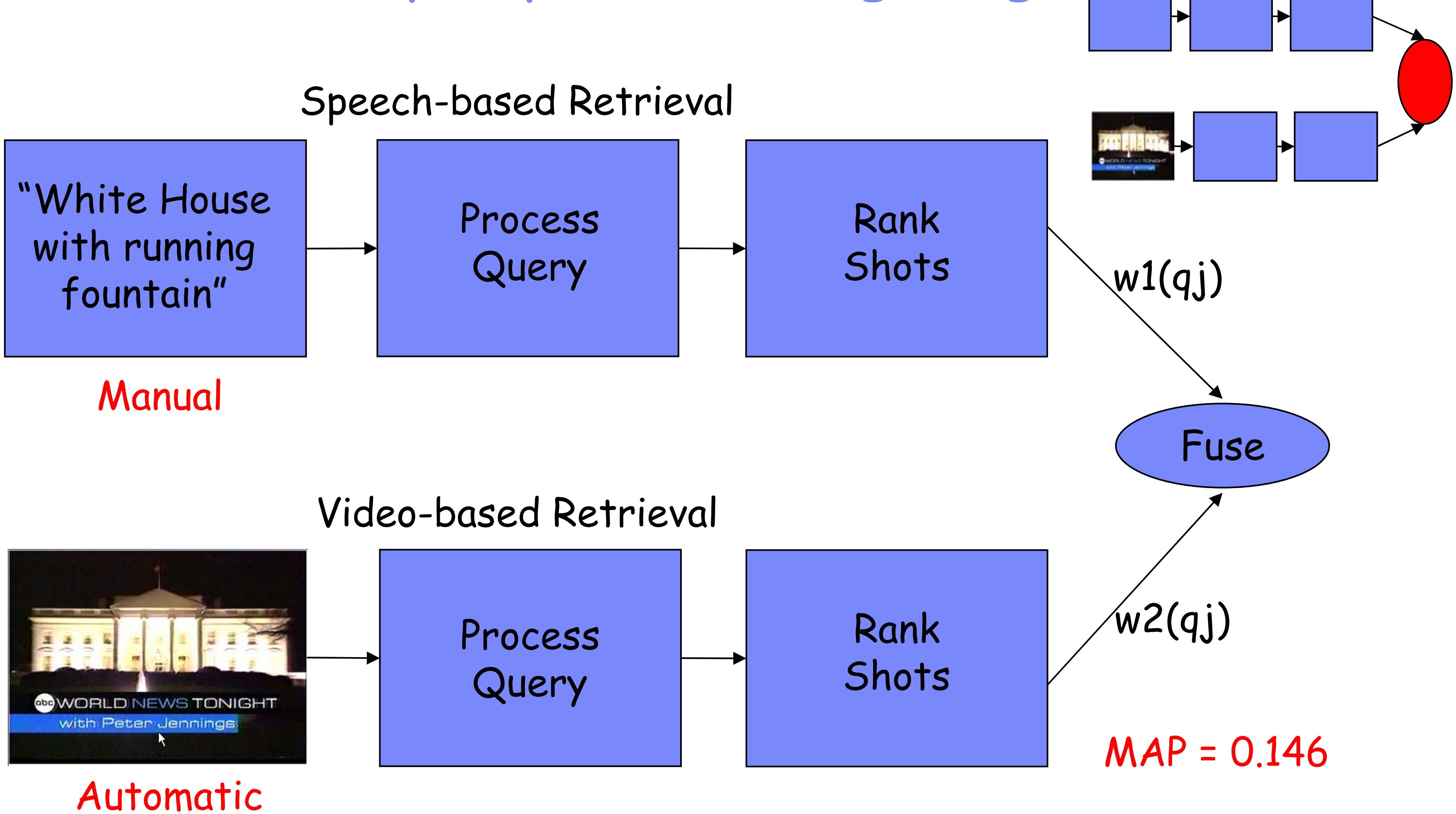


Original SDR AP = 0.23



Visually re-ranked AP = 0.27

Fusion II - Query dependent weighting



w1 and w2 are query dependent. w1+w2 = 1

Weights manually selected by the user based only on the query

Example: Baseball

- *"Find shots from behind the pitcher in a baseball game as he throws a ball that the batter swings at"*
- Manual SDR + automatic CBR
- Result of Manual Search on the Test set
- 60 of the top 100 are correct

Run	Average Precision
Best IBM	.39
Best non-IBM	.43
Average non-IBM	.125

Query Topics and Modality Performance

Query Types	Query Specificity			
	Generic		Specific/Named	
Find objects	Cats: Cup of coffee: Helicopters: Tanks:	********** ********** ********** **********	Sphinx: Tomb: Mercedes logo:	********** ********** **********
Find people	People diving: Urban people:	********** **********	Osama Bin Laden: Morgan Freeman: Pope John Paul II: Yasser Arafat: Mark Souder:	********** ********** ********** ********** **********
Find events	Rocket launch: Airplane take-off: Baseball pitch: Incoming train: Basketball hoop:	********** ********** ********** ********** **********	Dow Jones gain:	**********
Find scenes	Fires: Snow mountains: Aerial views: Roads with cars:	********** ********** ********** **********	White House:	**********

Legend:

***Speech

***Content

Better Modality Breakdown (# queries):

- Speech: 11
- Content: 9
- Either: 5

Conclusions

- Automatic video-MECBR is close to manual video-CBR
- Automatic SDR outperforms automatic/manual video-CBR
  - Speech modality better for 50-60% of the given query topics
- Multimodal runs outperformed unimodal runs
  - 20% improvement for manual runs, 40% for interactive runs
  - Improvement from last year's IBM performance
- System deficits:
  - Did not leverage annotators such as named entity detectors, face recognizers, text OCR, etc.
  - Most processing at shot keyframe level—hurts with long shots
- Late fusion approach: only explored limited schemes for system combination in the 15-minute limit
  - Query & data independent
  - Query dependent & data independent
  - Query and data dependent